**Requirements & Setup Of Spark NLP :**

Spark NLP is built on top of Apache Spark 3.x. For using Spark NLP you need:

* Java 8 and 11
* Apache Spark 3.3.x, 3.2.x, 3.1.x, 3.0.x

**Installation of Spark-nlp on windows:**

Follow the below steps to set up Spark NLP with Spark 3.2.3:

1.Download Adopt OpenJDK 1.8

* Make sure it is 64-bit
* Make sure you install it in the root of your main drive C:\java.
* During installation after changing the path, select setting Path.

2. Download the pre-compiled Hadoop binaries winutils.exe, hadoop.dll and put it in a folder called

C:\hadoop\bin from <https://github.com/cdarlint/winutils/tree/master/hadoop-3.2.0/bin>

Note : The version above is for Spark 3.2.3, which was built for Hadoop 3.2.0. You might have to change the hadoop version in the link, depending on which Spark version you are using.

3. Download Apache Spark 3.2.3 and extract it to C:\spark.

4. Set/add environment variables for HADOOP\_HOME to C:\hadoop and SPARK\_HOME to C:\spark.

5. Add %HADOOP\_HOME%\bin and %SPARK\_HOME%\bin to the PATH environment variable.

6. Install Microsoft Visual C++ 2010 Redistributed Package (x64).

7. Create folders C:\tmp and C:\tmp\hive

**Requisites for PySpark :**

We recommend using conda to manage your Python environment on Windows.

* Download Miniconda for Python 3.8
* See Quick Install on how to set up a conda environment with Spark NLP.
* The following environment variables need to be set:
  + PYSPARK\_PYTHON=python
  + Optionally, if you want to use the Jupyter Notebook runtime of Spark:
    - first install it in the environment with conda install notebook
    - then set PYSPARK\_DRIVER\_PYTHON=jupyter, PYSPARK\_DRIVER\_PYTHON\_OPTS=notebook
  + The environment variables can either be directly set in windows, or if only the conda env will be used, with conda env config vars set PYSPARK\_PYTHON=python. After setting the variable with conda, you need to deactivate and re-activate the environment.

Now you can use the downloaded binary by navigating to %SPARK\_HOME%\bin and running

Either create a conda env for python 3.6, install pyspark==3.3.1 spark-nlp numpy and use Jupyter/python console, or in the same conda env you can go to spark bin for pyspark –packages com.johnsnowlabs.nlp:spark-nlp\_2.12:4.3.1.

**Configuration :**

Python - 3.9

Apache spark - pyspark==3.2.1

spark-nlp==4.0.2

scala - 2.12

Spark NLP on CPU - Apache Spark - 3.0/3.1/3.2/3.3 - spark-nlp

-------------------------------------------------------------------------------------

# CPU

spark-shell --packages com.johnsnowlabs.nlp:spark-nlp\_2.12:4.0.2

pyspark --packages com.johnsnowlabs.nlp:spark-nlp\_2.12:4.0.2

spark-submit --packages com.johnsnowlabs.nlp:spark-nlp\_2.12:4.0.2

-----------------------------------------------------------------------------------